Supporting information for Omicron variant (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance

Jiahui Chen[1], Rui Wang[1], Nancy Gilby[2] and Guo-Wei Wei[1,3,4*]
[1] Department of Mathematics,
Michigan State University, MI 48824, USA.
[2] Spartan Innovations,
325 East Grand River Ave., Suite 355,
East Lansing, MI 48823 USA.
[3] Department of Electrical and Computer Engineering,
Michigan State University, MI 48824, USA.
[4] Department of Biochemistry and Molecular Biology,
Michigan State University, MI 48824, USA.

[*] Corresponding author. Email: weig@msu.edu

S1 Supplementary figures

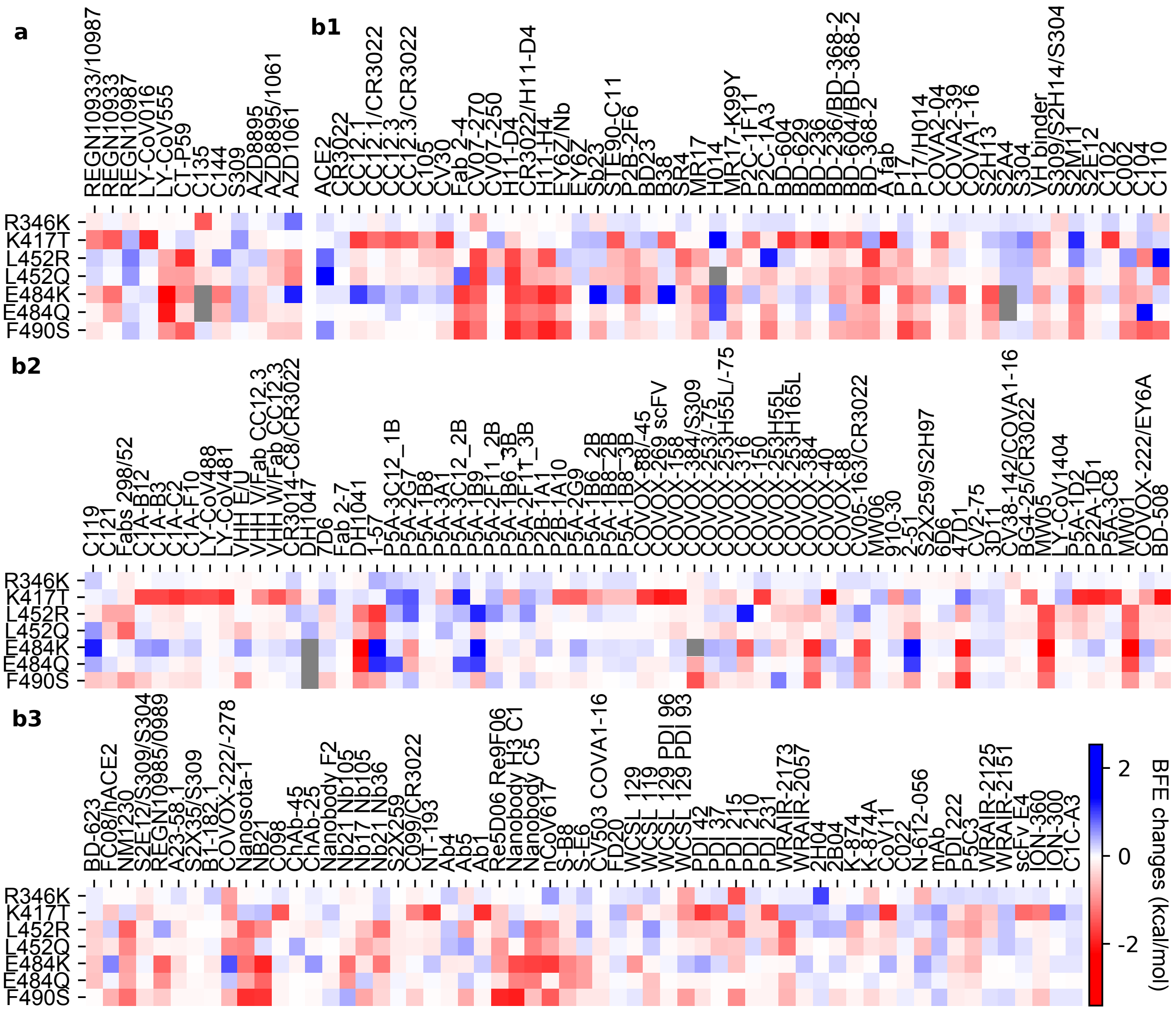

Figure S1: Illustration of BFE changes induced by RBD mutations R346K, K417T, L452R/Q, E484K/Q, F490S occurred in prevailing variants for 185 available antibody-RBD complexes and an ACE2-RBD complex. Positive changes strengthen the binding, while negative changes weaken the binding. **a** Heat map for 12 antibody and RBD complexes in various stages of drug development. Gray color stands for no predictions due to incomplete structures. **b1** Heat map for ACE2/antibody and RBD complexes. **b2** and **b3** Heat map for antibody and RBD complexes.

S2 Supplementary data

The Supplementary_Data.zip contains four files as listed in the following subsection.

S2.0.1 Disrupted antibodies

File antibodies_BFEs.csv shows the BFE changes of antibodies disrupted by Omicron mutations.

S2.0.2 List of antibodies

File antibodies.csv lists the Protein Data Bank (PDB) IDs for all of the 185 SARS-CoV-2 antibodies.

S3 Supplementary data pre-processing and feature generation methods

In this section, the workflow of the deep learning-based BFE change predictions of protein-protein interactions induced by mutations for the present SARS-CoV-2 variant analysis and prediction will be firstly introduced, which includes four steps as shown in Figure S2: (1) Data pre-processing; (2) training data preparation; (3) feature generations of protein-protein interaction complexes; and (4) prediction of protein-protein interactions by deep neural networks. Next, the validation of our machine learning-based model will be demonstrated, suggesting consistent and reliable results compared to the experimental deep mutations data.

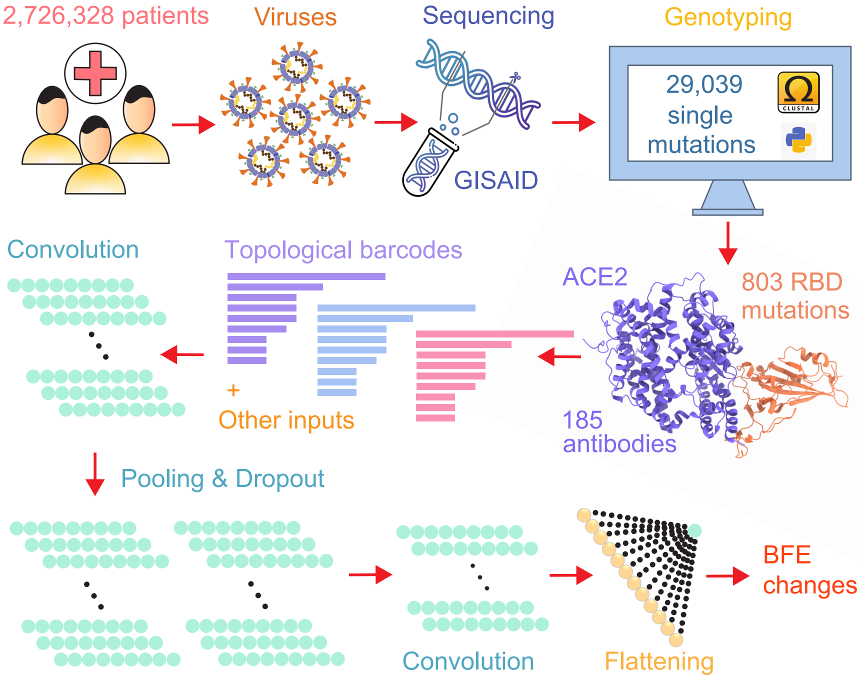

Figure S2: Illustration of genome sequence data pre-processing and BFE change predictions.

S3.1 Data pre-processing and SNP genotyping

The first step is to pre-process the original SARS-CoV-2 sequences data. In this step, a total of 2,726,328 complete SARS-CoV-2 genome sequences with high coverage and exact collection date are downloaded from the GISAID database [1] (https://www.gisaid.org/) as of November 20, 2021. Complete SARS-CoV-2 genome sequences are available from the GISAID database [1]. Next, the 2,726,328 complete SARS-CoV-2 genome sequences were rearranged according to the reference genome downloaded from the GenBank (NC_045512.2)[2], and multiple sequence alignment (MSA) is applied by using Cluster Omega with default parameters. Then, single nucleotide polymorphism (SNP) genotyping is applied to measure the genetic variations between different isolates of SARS-CoV-2 by analyzing the rearranged sequences [3, 4], which is of paramount importance for tracking the genotype changes during the pandemic. The SNP genotyping captures all of the differences between patients' sequences and the reference genome, which decodes a total of 29,039 unique single mutations from 2,726,328 complete SARS-CoV-2 genome sequences. Among them, 803 non-degenerate mutations on the S protein RBD (S protein residues from 329 to 530) are detected. In this work, the co-mutation analysis is more crucial than the unique single mutation analysis. Notably, the SARS-CoV-2 unique single mutations in the world are available at Mutation Tracker. The analysis of RBD mutations is available at Mutation Analyzer.

S3.2 Methods for BFE change predictions

In this section, the process of the machine learning-based BFE change predictions is introduced. Once the data pre-processing and SNP genotyping are carried out, we will firstly proceed with the training data preparation process, which plays a key role in reliability and accuracy. A library of 185 antibodies and RBD complexes, as well as an ACE2-RBD complex, are obtained from Protein Data Bank (PDB). RBD mutation-induced BFE changes of these complexes are evaluated by the following machine learning model. According to the emergency and the rapid change of RNA virus, it is rare to have massive experimental BFE change data of SARS-CoV-2, while, on the other hand, next-generation sequencing data is relatively easy to collect. In the training process, the dataset of BFE changes induced by mutations of the SKEMPI 2.0 dataset [5] is used as the basic training set, while next-generation sequencing datasets are added as assistant training sets. The SKEMPI 2.0 contains 7,085 single- and multi-point mutations and 4,169 elements of that in 319 different protein complexes used for the machine learning model training. The mutational scanning data consists of experimental data of the binding of ACE2 and RBD induced mutations on ACE2[6] and RBD[7, 8], and the binding of CTC-445.2 and RBD with mutations on both protein[8].

Next, the feature generations of protein-protein interaction complexes are performed. The element-specific algebraic topological analysis on complex structures is implemented to generate topological bar codes [9–12]. In addition, biochemistry and biophysics features such as Coulomb interactions, surface areas, electrostatics, et al., are combined with topological features [13]. The detailed information about the topology-based models will be demonstrated in subsection S3.3. Lastly, deep neural networks for SARS-CoV-2 are constructed for the BFE change prediction of protein-protein interactions [9]. The detailed descriptions of dataset and machine learning model are found in the literature [9, 14, 15] and are available at TopNetmAb.

S3.3 Feature generation for machine learning model

S3.3.1 Topology features

Among all features generated for machine learning prediction, the application of topology theory makes the model to a whole new level. Those summarized as other inputs are called as auxiliary features and are described in Section S3.3.2 and S3.3.3. In this section, a brief introduction about the theory of topology will be discussed. Algebraic topology [10, 11] has achieved tremendous success in many fields including biochemical and biophysical properties[12]. Special treatment should be implemented for biology applications to describe element types and amino acids in poly-peptide mathematically, which element-specific and site-specific persistent homology [14, 16]. To construct the algebraic topological features on protein-protein interaction model, a series of element subsets for complex structures should be defined, which considers atoms from the mutation sites, atoms in the neighborhood of the mutation site within a certain distance, atoms from antibody binding site, atoms from antigen binding site, and atoms in the system that belong to type of {C, N, O}, $\mathcal{A}_{\text{ele}}(\text{E})$. Under the element/site-specific construction, simplicial complexes is constructed on point clouds formed by atoms. For example, a set of independent $k+1$ points is from one element/site-specific set $U = \{u_0, u_1, ..., u_k\}$. The k-simplex σ is a convex hull of $k+1$ independent points U, which is a convex combination of independent points. For example, a 0-simplex is a point and a 1-simplex is an edge. Thus, a m-face of the k-simplex with $m+1$ vertices forms a convex hull in a lower dimension $m < k$ and is a subset of the $k+1$ vertices of a k-simplex, so that a sum of all its $(k-1)$–faces is the boundary of a k–simplex σ as

$$\partial_k \sigma = \sum_{i=1}^{k} (-1)^i \langle u_0, ..., \hat{u}_i, ..., u_k \rangle, \tag{1}$$

where $\langle u_0, ..., \hat{u}_i, ..., u_k \rangle$ consists of all vertices of σ excluding u_i. The collection of finitely many simplices is a simplicial complex. In the model, the Vietoris-Rips (VR) complex (if and only if $\mathbb{B}(u_{i_j}, r) \cap \mathbb{B}(u_{i_{j'}}, r) \neq \emptyset$ for $j, j' \in [0, k]$) is for dimension 0 topology, and alpha complex (if and only if $\cap_{u_{i_j} \in \sigma} \mathbb{B}(u_{i_j}, r) \neq \emptyset$) is for point cloud of dimensions 1 and 2 topology [12].

The k-chain c_k of a simplicial complex K is a formal sum of the k-simplices in K, which is $c_k = \sum \alpha_i \sigma_i$, where α_i is coefficients and is chosen to be $\mathbb{Z}_2$. Thus, the boundary operator on a k-chain c_k is

$$\partial_k c_k = \sum \alpha_i \partial_k \sigma_i, \tag{2}$$

such that $\partial_k : C_k \to C_{k-1}$ and follows from that boundaries are boundaryless $\partial_{k-1}\partial_k = \emptyset$. A chain complex is

$$\cdots \xrightarrow{\partial_{i+1}} C_i(K) \xrightarrow{\partial_i} C_{i-1}(K) \xrightarrow{\partial_{i-1}} \cdots \xrightarrow{\partial_2} C_1(K) \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0, \tag{3}$$

as a sequence of complexes by boundary maps. Therefore, the Betti numbers are given as the ranks of kth homology group H_k as $\beta_k = \text{rank}(H_k)$, where $H_k = Z_k/B_k$, k-cycle group Z_k and the k-boundary group B_k. The Betti numbers are the key for topological features, where β_0 gives the number of connected components, such as number of atoms, β_1 is the number of cycles in the complex structure, and β_2 illustrates the number of cavities. This presents abstract properties of the 3D structure.

Finally, only one simplicial complex couldn't give the whole picture of the protein-protein interaction structure. A filtration of a topology space is needed to extract more properties. A filtration is a nested sequence such that

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K. \tag{4}$$

Each element of the sequence could generate the Betti numbers $\{\beta_0, \beta_1, \beta_2\}$ and consequentially, a series of Betti numbers in three dimensions is constructed and applied to be the topological fingerprints in Figure S2.

S3.3.2 Residue-level features

Mutation site neighborhood amino acid composition Neighbor residues are the residues within 10 Å of the mutation site. Distances between residues are calculated based on residue C_α atoms. Six categories of amino acid residues are counted, which are hydrophobic, polar, positively charged, negatively charged, special cases, and pharmacophore changes. The count and percentage of the 6 amino acid groups in the neighbor site are regrading as the environment composition features of the mutation site. The sum, average, and variance of residue volumes, surface areas, weights, and hydropathy scores are used but only the sum of charges is included.

pKa shifts The pKa values are calculated by the PROPKA software [17], namely the values of 7 ionizable amino acids, namely, ASP, GLU, ARG, LYS, HIS, CYS, and TYR. The maximum, minimum, sum, the sum of absolute values, and the minimum of the absolute value of total pKa shifts are calculated. We also consider the difference of pKa values between a wild type and its mutant. Additionally, the sum and the sum of the absolute value of pKa shifts based on ionizable amino acid groups are included.

Position-specific scoring matrix (PSSM) Features are computed from the conservation scores in the position-specific scoring matrix of the mutation site for the wild type and the mutant as well as their difference. The conservation scores are generated by PSI-BLAST [18].

Secondary structure The SPIDER2 software is used to compute the probability scores for residue torsion angle and residues being in a coil, alpha helix, and beta strand based on the sequences for the wild type and the mutant [19].

S3.3.3 Atom-level features

Seven groups of atom types, including C, N, O, S, H, all heavy atoms, and all atoms, are considered when generating the element-type features. Meanwhile, other three atom types, i.e., mutation site atoms, all heavy atoms, and all atoms, are used when generating the general atom-level features.

Surface areas Atom-level solvent excluded surface areas are computed by ESES [20].

Partial changes Partial change of each atom is generated by pdb2pqr software [21] using the Amber force field [22] for wild type and CHARMM force field [23] for mutant. The sum of the partial charges and the sum of absolute values of partial charges for each atomic group are collected.

Atomic pairwise interaction interactions Coulomb energy of the ith single atom is calculated as the sum of pairwise coulomb energy with every other atom as

$$C_i = \sum_{j,j\neq i} k_e \frac{q_i q_j}{r_{ij}}, \tag{5}$$

where k_e is the Coulomb's constant, r_{ij} is the distance of ith atom to jth atom, and q_i is the charge of ith atom. The van der Waals energy of the ith atom is modeled as the sum of pairwise Lennard-Jones potentials with other atoms as

$$V_i = \sum_{j,j\neq i} \epsilon \Big[\big(\frac{r_i + r_j}{r_{ij}} \big)^{12} - 2 \big(\frac{r_i + r_j}{r_{ij}} \big)^{6} \Big], \tag{6}$$

where ϵ is the depth of the potential well, and r_i is van der Waals radii.

In atomic pairwise interaction, 5 groups (C, N, O, S, and all heavy atoms) are counted both for Coulomb interaction energy and van der Waals interaction energy.

Electrostatic solvation free energy Electrostatic solvation free energy of each atom is calculated using the Poisson-Boltzmann equation via MIBPB [24] and are summed up by atom groups.

S4 Supplementary machine learning methods

The topology-based network model for BFE change predictions induced mutations on SARS-CoV-2 studying applies a deep neural network structure. Similar approaches have been widely implemented in the energy prediction of protein-ligand binding[25] and protein-protein interactions[16]. The neural network method maps an input feature layer to output layer and mimics biological brains for solving problems where numerous neuron units are involved and weights of neurons are updated by backpropagation methods. To make more complicated structure in order to extract abstract properties, more layers and more neurons in each layer can be constructed. In the training process, optimization methods are applied to avoid overfitting issue, such as dropout methods[26] that a partial of computed neurons of each layer is dropped. For the model cross validations, the Pearson correlation of 10-fold cross-validation is 0.864, and the root mean square error is 1.019 kcal/mol.

S4.1 Deep learning algorithms

A deep neural network is a neural network method with multi-layers (hidden layers) of neurons between the input and output layers. In each layer, the single neuron gets fully connected with the neurons in the next layer. It should preserve the consistency of all labels when applying the model for mutation-induced BFE change predictions. The loss function is constructed as follows:

$$\underset{W,b}{\operatorname{argmin}} L(W,b) = \underset{W,b}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \big(y_i - f(x_i; \{W,b\}) \big)^2 + \lambda \|W\|^2 \tag{7}$$

where N is the number of samples, f is a function of the feature vector x_i parameterized by a weight vector W and bias term b, and λ represents a penalty constant.

S4.2 Optimization

The backpropagation is applied to evaluate the loss function starting from the output layer and propagates backward through the network structure to update the weight vector W and bias term b. Since gradient calculation is required, therefore, we apply the stochastic gradient descent method with momentum, which only evaluates a small part of training data and can be considered as calculating exponentially weighted averages, which is given as

$$\begin{aligned} V_i &= \beta V_{i-1} + \eta \nabla_{W_i} L(W_i, b_i) \\ W_{i+1} &= W_i - V_i, \end{aligned} \tag{8}$$

where W_i is the parameters in the network, $L(W_i, b_i)$ is the objective function, η is the learning rate, X and y are the input and target of the training set, and $\beta \in [0, 1]$ is a scalar coefficient for the momentum term. The momentum term involved accelerates the converging speed.

S5 Supplementary validation

In the main content, we briefly summarized validations of our machine learning predictions and experimental data. For large quantitative validations, we compared the BFE change prediction for mutations on S protein RBD to the experimental deep mutational enrichment data on RBD binding to human ACE2 and CTC-445.2 induced by RBD mutations [8, 9, 13]. To make these validations, we eliminated the experimental deep mutational enrichment data of RBD binding to human ACE2 and CTC-445.2 from the training sets and set them as testing sets, which have 1539 and 1500 samples, respectively. In the validation of RBD and CTC-445.2 complex, there is a very high correlation between the enrichment data and machine learning predictions, as well as the validation of RBD binding to ACE2, with Pearson correlations are 0.69 and 0.70, respectively. The deep mutational enrichment data can give a proportional descriptor of the affinity strength of protein-protein interactions induced by mutations. The machine learning methods, however, give stable and equalized evaluations, while experimental data might be different dramatically due to conditions and environments.

In addition, we compared our machine learning results with other experimental data, which are escape fraction, pseudovirus infection changes, and IC_{50} fold changes [9]. In the comparison of 35 cases to experimental escape fractions on RBD binding to clinical trial antibodies induced by emerging mutations, our machine learning predictions have a Pearson correlation of 0.80. Especially, those high escaping mutations E484K and E484Q on LY-CoV555, and mutations K417T and K417N on LY-CoV016, are indicated by both our predictions and the experimental data [9]. We also use the pattern comparisons of our prediction to experimental data. Lastly, we collected experimental data from different literature [27–30]. According to variations from different research groups, they were summarized in increasing/decreasing patterns of emerging variant (including co-mutations) impacts on antibody therapies in clinical trials. In total, there are 20 pattern comparisons with an excellent agreement between various experimental data and our predictions, except for a minor discrepancy [9].

References

(1) Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **2017**, *22*, 30494.

(2) Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y., et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269.

(3) Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* **2020**, *112*, 3588–3596.

(4) Kim, S.; Misra, A. SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* **2007**, *9*, 289–320.

(5) Jankauskaitė, J.; Jiménez-Garcıa, B.; Dapkūnas, J.; Fernández-Recio, J.; Moal, I. H. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **2019**, *35*, 462–469.

(6) Procko, E. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *BioRxiv* **2020**.

(7) Starr, T. N.; Greaney, A. J.; Hilton, S. K.; Ellis, D.; Crawford, K. H.; Dingens, A. S.; Navarro, M. J.; Bowen, J. E.; Tortorici, M. A.; Walls, A. C., et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **2020**, *182*, 1295–1310.

(8) Linsky, T. W.; Vergara, R.; Codina, N.; Nelson, J. W.; Walker, M. J.; Su, W.; Barnes, C. O.; Hsiang, T.-Y.; Esser-Nobis, K.; Yu, K., et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science* **2020**, *370*, 1208–1214.

(9) Chen, J.; Gao, K.; Wang, R.; Wei, G.-W. Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies. *J. Mol. Biol.* **2021**, *433*.

(10) Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **2009**, *46*, 255–308.

(11) Edelsbrunner, H.; Letscher, D.; Zomorodian, A. In *Proceedings 41st annual symposium on foundations of computer science*, 2000, pp 454–463.

(12) Xia, K.; Wei, G.-W. Persistent homology analysis of protein structure, flexibility, and folding. *Int J Numer Method Biomed Eng* **2014**, *30*, 814–844.

(13) Chen, J.; Gao, K.; Wang, R.; Wei, G.-W. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chem. Sci.* **2021**, *12*, 6929–6948.

(14) Chen, J.; Wang, R.; Wang, M.; Wei, G.-W. Mutations strengthened SARS-CoV-2 infectivity. *J. Mol. Biol.* **2020**, *432*, 5212–5226.

(15) Wang, R.; Hozumi, Y.; Yin, C.; Wei, G.-W. Mutations on COVID-19 diagnostic targets. *Genomics* **2020**, *112*, 5204–5213.

(16) Wang, M.; Cang, Z.; Wei, G.-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nat. Mach. Intell.* **2020**, *2*, 116–123.

(17) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics* **2008**, *73*, 765–783.

(18) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(19) Yang, Y.; Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y. In *Prediction of protein secondary structure*; Springer: 2017, pp 55–63.

(20) Liu, B.; Wang, B.; Zhao, R.; Tong, Y.; Wei, G.-W. Eses: software for e ulerian solvent excluded surface, 2017.

(21) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.

(22) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W., et al. *Amber 10*; tech. rep.; University of California, 2008.

(23) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem* **2009**, *30*, 1545–1614.

(24) Chen, D.; Chen, Z.; Chen, C.; Geng, W.; Wei, G.-W. MIBPB: a software package for electrostatic analysis. *J. Comput. Chem* **2011**, *32*, 756–770.

(25) Nguyen, D. D.; Wei, G.-W. AGL-Score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J Chem Inf Model* **2019**, *59*, 3291–3304.

(26) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* **2014**, *15*, 1929–1958.

(27) FACT SHEET FOR HEALTH CARE PROVIDERS EMERGENCY USE AUTHORIZATION (EUA) OF REGEN-COV (fda.gov).

(28) Weisblum, Y.; Schmidt, F.; Zhang, F.; DaSilva, J.; Poston, D.; Lorenzi, J. C.; Muecksch, F.; Rutkowska, M.; Hoffmann, H.-H.; Michailidis, E., et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* **2020**, *9*, e61312.

(29) Wang, P.; Nair, M. S.; Liu, L.; Iketani, S.; Luo, Y.; Guo, Y.; Wang, M.; Yu, J.; Zhang, B.; Kwong, P. D., et al. Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. *Nature* **2021**, *10*.

(30) Planas, D.; Veyer, D.; Baidaliuk, A.; Staropoli, I.; Guivel-Benhassine, F.; Rajah, M. M.; Planchais, C.; Porrot, F.; Robillard, N.; Puech, J., et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **2021**, 1–7.